DOCUMENT RESUME

ED 078 602 EA 005 347

AUTHOR	Raizen, Senta A.
TITLE	Evaluating the National Institute of Education.
INSTITUTION	Rand Corp., Santa Monica, Calif.
REPORT NO	Rand-P-4942
PUB DATE	Feb 73
NOTE	29p.
AVAILABLE FROM	Publications Department, Rand, 1700 Main Street, Santa Monica, California 90406 (Order No. P-4942, $2.00)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Educational Development; *Educational Research; Evaluation; *Evaluation Criteria; *Evaluation Methods; *Federal Programs; Research Criteria
IDENTIFIERS	*National Institute of Education; NIE

ABSTRACT

This paper begins by examining the context within which the NIE is likely to be evaluated, such as what it will be expected to achieve, who will assess its achievements, and why and when assessments will be made. The second part of the paper deals with suggested evaluation dimensions of (1) the technical quality of the R&D supported by NIE, (2) the choice of questions or problems being addressed, (3) the effectiveness of program output, and (4) distribution of funds and other benefits. For each of these dimensions, some criteria and evaluation methods are given. The last section discusses briefly how evaluation results could be used to improve the performance of NIE by application to resource allocation, management procedures, and organization. (Author)

FILMED FROM BEST AVAILABLE COPY

ED 078602

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Rand Corp.
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

EVALUATING THE NATIONAL INSTITUTE OF EDUCATION

Senta A. Raizen

February 1973

EA 005 347

P-4942

FILMED FROM BEST AVAILABLE COPY

Any views expressed in this paper are those of the authors. They should not be interpreted as reflecting the views of The Rand Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The Rand Corporation as a courtesy to members of its staff.

EVALUATING THE NATIONAL INSTITUTE OF EDUCATION

Senta A. Raizen*
The Rand Corporation, Santa Monica, California

The introduction in 1965 of the Planning-Programming-Budgeting system into government has brought in its wake increasing demands for evaluating the effectiveness of government programs. As experience with assessment of social programs, and particularly educational experimentations, has accumulated, knowledgeable researchers have come to advocate that evaluation be made an integral part of program development (Light and Smith, 1970; Smith and Bissell, 1970). But to my knowledge, it is unique for a government R&D agency to consider at its inception by what standards its accomplishments should be judged. Perhaps in the case of the National Institute of Education (NIE), which was created to solve educational problems, this early concern with evaluation is an expression of the current disenchantment with R&D, or of the erosion of the formerly deep-rooted American faith in education as the solution to most social problems. Indeed, questions have been raised as to the impact that a government R&D agency can have on education, in view of our limited knowledge about education and R&D (Cohen, 1972). Further, some critics do not hold out much hope for evaluation as a consequential means of influencing policy or practice (Fox, 1967). Since the creation of a new government agency such as the NIE is itself a form of social action, however, attempts to assess its effectiveness will inevitably be made. If the NIE can guide these attempts by developing legitimate standards for its evaluation so that results will be both useful and actually used, one of the NIE's first successes may be to provide an acceptable pattern for other government agencies.

---

*An earlier version of this paper was presented at a Symposium on Educational Research and Development sponsored by the NIE in Washington, D.C., on December 11, 1972.

Rossi and Williams (1972) note a number of problems and risks in developing and using evaluative results: conceptual and methodological problems, scarcity of competent people, and bureaucratic and political risks. Bureaucratic and political impediments are likely to operate particularly strongly in the case of evaluating overall performance of an agency. This is not only because, "...neither social service delivery systems nor government programs are organized to generate information about their effectiveness" (Rivlin, 1971, p. 64), but also because, as the new director of the NIE himself stated in a paper appraising the evaluation of federal manpower programs (Glennan, 1969, p. 45), "(1) Most programs and most agencies are reluctant to be evaluated; (2) if they must be evaluated, they will seek to find evaluation designs that have the greatest probability of supporting the status quo."

This paper is an effort to clarify the problem of evaluating the NIE and begin the process of developing satisfactory performance criteria. It assumes that R&D carried out by a federal agency can contribute to education, and that it is therefore in the interest of all -- the government, the R&D communities, the education professions, students and parents, and taxpayers -- to see the agency succeed. Unfortunately, the paper has had to be prepared in the absence of NIE-generated programs and of organizational structure for the new agency, thus limiting the discussion of evaluation methods and criteria to rather general and abstract suggestions. These need to be developed in greater detail as agency plans and organization are formulated.

I. THE CONTEXT FOR EVALUATION OF THE NIE

To understand just how difficult the task is, we must take a look at the context within which the NIE is going to operate. The old five "w's" of the newspaper reporter -- *who*, *what*, *why*, *when*, *where* -- can help us define this context.

Let me change the order slightly and deal with the *what* first. This is essentially a question of defining the objectives of the NIE. The legislative charter of the NIE is not much help since it is framed so broadly that it merely transforms the overall question of whether, and how much, the NIE is improving education into four questions, namely

whether, and how much, the NIE is:

"(A) helping to solve or to alleviate the problems of, and achieve the objectives of American education;

(B) advancing the practice of education, as an art, science, and profession;

(C) ...strengthening...the scientific and technological foundations of education; and

(D) building an effective educational research and development system."*

In order to *evaluate* anything, as the word implies, one must know clearly what is being valued, that is, one must define the desired directions of change. Lord Rothschild (1972) advocates that the client of R&D -- either the government agency or the ultimate users -- decide specific program objectives. It would appear, therefore, that one of the first tasks for the NIE is to translate the general goals in its legislation into operational objectives that reflect consumer needs. For the NIE, this is a rather more complex undertaking than for an agency like the NIH, on which the NIE is superficially modeled. There are clear consensus goals for R&D in health: curing cancer, reducing the incidence of dental caries, eliminating stroke and heart disease. While the choice of strategies and resource allocations for R&D to attain these goals may often be difficult, at least they are undisputed social goods and clearly perceived. However, as we move into the domains of mental health and human development where goals are less easily defined, we find the cognizant agencies having greater difficulty justifying their programs. In the case of education, there appear indeed to be consensus goals, at least at the level of public rhetoric: ability to get a "good" job -- preferably one yielding financial success,** learning to get along with others,** achieving some acceptable level of reading and other cognitive skills by the end of high school, escaping poverty, making the schools work better for the children of the poor. But by now we know that all these goals have

---

*These four missions are quoted from *Education Amendments of 1972*.

**Obtaining a better job (44 percent), getting along better with people (43 percent), and financial success (38 percent) were the top goals for education of their children chosen by adults in a recent Gallup Poll reported in *Phi Delta Kappan*, September 197[illegible], p. 33.

one of two (sometimes both) characteristics: They are not equally valued by different client groups and therefore often in competition for resources with each other and with other highly valued educational goals, or they cannot be attained solely or even primarily through education. For example, investing the sizable resources necessary to make the bottom ten percent of students reach reading norms would either require withdrawing resources from other instructional areas now considered important or conflict directly with another current goal -- holding the line on steeply rising educational expenditures. Attaining a good job requires a set of attitudes, abilities, and credentials to which education can contribute, but it also requires that there be enough good jobs to go around, a function of the labor market rather than educational processes.

Thus, the question of *what* becomes one of defining important and relevant problems, relevant in the sense that they actually fall within the domain of education and are amenable to R&D approaches. The NIE can itself contribute to the validity of judgments about its programs by phrasing its objectives to imply appropriate rather than inappropriate criteria. To elaborate the job example, reduction of unemployment is an inappropriate criterion for R&D in education, but additional understanding of who is unemployed because of lack of skills, and the degree of success of new educational systems that deliver the needed skills are both relevant to assessing a program concerned with R&D in career education.

An equally important concern is: *who* should evaluate the NIE? Or, perhaps more realistically, who will evaluate the NIE? This is a complicated question that can be answered by a simple declarative sentence: The NIE is an agency of the federal government in the Department of Health, Education, and Welfare charged with carrying out R&D in education. Each of the nouns in this sentence corresponds to a set of evaluators. Any agency of the federal government will be judged by the Executive Office (currently that means largely the OMB) in the context of the President's annual budget formulation, and by the cognizant committees of Congress in the context of legislation (authorization committees) and financing (appropriation committees). Any component of HEW must also account

to the Secretary, the Assistant Secretary for Planning and Evaluation, and -- in the case of an education agency -- the Assistant Secretary for Education. An R&D agency will be judged by the R&D communities that are or would like to be its client groups; an agency created to improve education will receive critical attention from all those who have a stake in that enterprise.

To appreciate the special difficulties faced by the NIE in regard to the last two groups, a comparison with the NIH is again illuminating. The R&D communities that interact with the NIH are well-defined and share a common set of belief structures and interests, not only because they are drawn from the disciplinary bases of the biomedical sciences, but because the first director of the NIH made it his prime responsibility to establish the desired relationships. The case of the NIE is quite different; in addition to the established educational research community, researchers from many different disciplines and applied fields -- from the natural and behavioral sciences to the humanities, from operations research to communications technology -- can and do claim that they can contribute to, and therefore assess, R&D in education.* These disparate groups hardly understand each other's languages; instead of sharing a common outlook, they are ignorant of each other or, if brought into contact, are often sharply at odds in defining R&D problems, in preferred R&D styles, and in assessing outcomes. As to those who have a stake in education, the NIE is faced with two facts: powerful and vocal professional organizations (some two million strong) who consider themselves in an expert position to judge the effectiveness of educational R&D, and -- in great contrast to medicine -- the widespread belief on the part of the ultimate consumer (student, parent, employer, taxpayer) that he knows quite as much about education, having gone through it, as the professional.

As one considers the kinds of questions these different overseers are likely to ask, one comes to the *why*, the purpose of evaluation.

---

*Insofar as the NIH is having to concern itself increasingly with delivery of health services, it will have to involve a wider spectrum of performers, and its problems will be more like those of the NIE.

Levine and Williams (1971) note two such purposes: to affect resource allocation, and to improve R&D strategies. The governmental overseers are likely to be most concerned with the first purpose; the R&D communities, with both allocations and strategies; and the education professionals and consumers, with the eventual results of allocations and strategies. Resource allocations are usually the result of many considerations, however, and objective assessment of the benefits of a particular program or agency as compared to others with claims on the federal dollar often play only a minor part. Evaluation cannot, and should not, serve as a substitute for good judgment. Nevertheless, the HEW leadership and OMB will expect evaluation to produce information on the potential importance of each NIE program, potential payoff, and likelihood of success. Congress will have similar concerns, and in addition be sensitive to geographic and institutional distributions of funds and second-order benefits. The NIE and its advisers would do well to construct ahead of time some evaluation criteria responsive to such questions; I shall try to suggest a few later in this paper.

The various R&D communities will (whether asked or not) assess the quality of the R&D output, with implications for improvements in R&D strategy; a concern with who receives funding for what purposes will hardly be divorced from this assessment, but the criteria for appropriate distribution will no doubt be quite different from those of Congress. The judgment of consumers is likely to be influenced by governmental and R&D performer groups insofar as their evaluations receive public notice -- though the influence often may take opposite directions for different consumers. But consumers will react with much greater intensity to programs having a direct impact on them, either as practitioners or as recipients of education. This reaction can be justified, as in the case of frustration with performance of inner-city schools in the face of ESEA Title I efforts, or capricious, as witness the furor over the "new math." Although the NIE will not be able to discount the unsolicited judgments of clients and consumers, these are not likely always to provide the optimal input for improving

its R&D strategies and operations.* There is another group of observers, the Advisory Council, that should be specifically charged with the responsibility of synthesizing judgments from all the evaluating groups -- self-appointed and solicited -- in order to make its own assessments and provide feedback to the NIE on needed changes to improve its performance.

The *when* is a critical problem for the NIE. Federal resource decisions are made in the course of the annual budget cycle; the present climate for educational R&D is not likely to permit growth or perhaps even stability without evidence as to achievement for dollars invested even within the first couple of years, unreasonable as that may seem given the difficulty of some of the problems and the long-range effects of most educational interventions. This implies that, no matter what other considerations go into choice of programs, there must be some activities designed to yield short-term successes, and some which are convincing demonstrations that progress is being made toward solving some difficult problems. Again, it will be necessary to spell out appropriate indicators for such successes and demonstrations, so that rational assessment can inform the decisions that are going to be made in any case by the executive branch and Congress.

One would hope for more leadtime for judgments from the R&D communities, education professionals, and ultimate consumers on the efficacy of a new agency's programs and operations, but here also history has predisposed many of us (for we all fall into one of these three groups) toward impatience, if not skepticism. Signals as to its competence will have to be given early if the NIE is to avoid a premature -- and negative -- assessment of "more of the same." Thus, the NIE faces severe time constraints, yet planning for long-range evaluation may be as important for its future as concern for immediate survival. Some of the most significant contributions of R&D to education are likely to be efforts resulting in the design of improved products, practices, and perhaps entire new systems of delivery for education, and such efforts may well consume five to ten years, with valid assessment stretching even beyond.

---

*The NIE might, however, develop consumer-oriented evaluation procedures for products developed under its sponsorship to clarify the purposes and appropriate applications of those products.

Therefore, while the NIE will have no choice about short-range accountability, it must explicitly build toward a demonstrable record of achievement measured on a time scale appropriate to design efforts.*

In journalistic practice, the *where* tends to come at the bottom of the list as providing the least important bit of information. In the present context, it might be considered synonymous with *who* if we interpret it as meaning where assessment questions will be asked. I wish, however, to consider instead *where* they might be answered, or better, *how* they might be answered. The remainder of this paper will, therefore, deal with the where-how of evaluating the NIE in the climate of the existing constraints.

II. THE DIMENSIONS OF EVALUATION

The dimensions of evaluation are implicit in who is rendering judgment with what purpose. They can be subsumed under four general headings: technical quality of the R&D, choice of the questions or problems being addressed, effectiveness of program output, and distribution of funds and of second-order benefits. Each of these dimensions has associated with it a series of questions that can help us define relevant criteria and perhaps even some appropriate methodologies.

Technical Quality**

The caliber of the R&D supported by the NIE is of most direct concern to the R&D communities, although it will, in the long run, affect the judgments of other groups as well, as quality begins to impact on the agency's ability to address problems these groups perceive to be important. Some questions useful in structuring any assessment of R&D quality are:

- o What are the fields (and subfields) of activity?
- o On what basis are they selected?

---

*Iterative engineering characterizes successful design. Whether the design is to be for a hardware system such as a moon launch or for a service system such as design and implementation of an innovative curriculum, the time span needed tends to be measured in decades rather than in single years.

**This section draws on some unpublished work by John Wirt of The Rand Corporation, who kindly made it available to me.

- o What are the objectives in each field and subfield?
- o What styles of R&D (research, policy analysis, development/ design, experimentation, evaluation) are being supported?
- o Is the mix of styles appropriate to the objectives in each field?
- o What is the quality of the performers being supported?
- o What is the mix of performers?
- o Is this mix appropriate to the objectives in each field?
- o What contributions are being made to the knowledge base in each field?

Of the four different aspects of evaluating a mission-oriented R&D agency considered in this paper, the methodology for quality assessment is probably the best developed. Criteria for choosing fields and subfields have been established in such recent examinations as *Priorities for Space Research, 1971-1980* (1971), and *Physics in Perspective: Recommendations and Program Emphases* (1972). They generally include intrinsic, extrinsic, and collateral criteria. Intrinsic criteria measure inherent quality: "ripeness" of the field, availability of new techniques, recent discoveries that have posed new significant questions, prospects of opening up further areas of inquiry, propensity of the field to attract able researchers. Extrinsic criteria are concerned with contributions to other fields, to policy, to progress in practical applications to social goals. Collateral criteria deal with coherence of R&D activities within the agency, coherence in the context of overall activities in a field, consistency and reliability of results, appropriate balance of R&D styles.

Relatively well-established practices for applying these criteria to R&D programs exist. The raw material for evaluation is aggregate information on proposals, performers, progress reports, final reports, and review information on R&D projects supported. The method usually involves some form of peer-group review, often through specially convened panels, sometimes via a two-tier system consisting of specialist subpanels and an overview panel.* For the NIE as a whole, the Advisory

*The NIE has itself applied this type of assessment to the programs of the Regional Laboratories and R&D Centers.

Council could function as the overview panel, although completely independent reviews should also take place to assure objectivity and credibility. In actual applications, the procedure often resembles an adversary model in "which there are claims and counterclaims, arguments and counter-arguments, and each side advanced by an advocate who attempts to make the best possible case for his position [*sic*]."* Guttentag (1971) points out that this is a quite appropriate model for evaluating programs in actual social contexts, but it requires advocates deeply versed in the case -- and that is, of course, the catch for the NIE. Criteria will be applied differently by educational researchers as opposed to behavioral scientists, by economists as opposed to computer scientists, and so on. Whose advice should the NIE elicit to improve its programs? Whose opinion will it have to take into account, whether the assessment was elicited or not? Whom will the other groups, particularly those that control NIE's resources, listen to on questions of quality of its programs? The responses that the NIE formulates to these questions -- the relationships that it chooses to build, as in the case of the NIH -- will play a major role in its development, perhaps even its survival. And there is no substitute for staff competence and judgment in shaping these relationships.

An auxiliary mechanism coming into more frequent use to help structure technical evaluations of R&D programs is the commissioning of state-of-the-art reviews. Such reviews can be considered the research component of evaluating an R&D program, for they investigate ("gather evidence" on) the content of each field, its strengths and weaknesses, the record of progress. While panel evaluation should take place periodically, perhaps annually or biennially, state-of-the-art reviews ought to be carried on

---

*Guttentag (1971) quoting from an unpublished paper by M. Levine. Cain and Hollister (1969) also discuss evaluation as "an attempt to raise the standards of what is admissible as evidence in a decision process that is inherently likely to remain adversary in nature. Higher standards of evaluation will lessen the role of 'hearsay' testimony in the decision process, but they are not meant to provide a hard and fast decision rule in and of themselves...if standards for the acceptance of evaluation results are viewed in terms of the 'rules of evidence' analogy, we can begin to move toward the judicious mix of rigor and pragmatism that is so badly needed in evaluation analysis."

continually, field by field, and each field should be reviewed every few years. The NIE itself should sponsor the reviews as one of its research activities. Staffing (whether by in-house researchers, outside consultants and grantees, or a combination) is critical; the individuals charged with conducting reviews of a field must be able to commit enough time, be technical experts, have wide-ranging interests in order to avoid biases, and have highly developed critical and analytical faculties; the actual authors of resulting papers must also have lucid writing styles. Some questions that can be addressed by state-of-the-art reviews include:*

- o What are the principal findings and results in the field being examined? In each subfield?
- o What are the principal non-findings?
- o What is the technical reliability of results? Are achievements repeatable? Consistent?
- o What R&D problems are currently receiving the most attention? Why?
- o What problems are important but not being worked on? Why not?
- o What are the principal impediments to more rapid technical progress? Lack of data? Lack of theory? Lack of facilities or appropriate settings? Lack of instrumentation?
- o Who are the major contributors to the field?

Findings developed by such reviews of fields and subfields should be published in professional journals and other media, for, if well done, they can provide milestones not only for NIE planning and evaluation, but for the wider community of researchers, professionals, and interested laymen. This would be a useful service for the NIE to perform, quite apart from supplying input for assessing the progress being made through its support of R&D.

Choice of Questions or Problems

The NIE might receive high marks for the technical quality of the R&D it supports, and yet be condemned on the basis of not coming to

---

*A detailed list, much of which is applicable, can be found in Appendix C of *Physics in Perspective* (1972).

grips with the really important problems of education as commanded by its charter. One observer (Timpane, 1970, p. 565) comes to rather pessimistic conclusions as to the ability of R&D to deal with priority problems: "If a problem area proposed for experimentation is unpopular and/or unimportant, experimentation should not and/or will not be done; but if it is popular and important, action will not wait for experimentation.... In the competition for funds, short-term attention to action demands is likely to offer greater promise of political reward than research." In the face of such political exigencies, will the NIE be able to address important problems? And in the absence of clear consensual goals and sufficient understanding of problems to allow parceling out the educational components, how can one assess whether the problems the NIE does select to work on are the right ones? These questions will not be satisfied by an evaluation of the kind just described, which is concerned with research and technical problems, for clearly the word "problems" in the legislation and in the view of most of the NIE's overseers (excepting only some components of the R&D community) carries a quite different meaning. It refers to the publicly perceived educational problems, for example, the failure of the schools to teach reading, and not uncommonly even includes non-educational problems thought to be solvable through education, such as drug abuse or environmental deterioration. Insofar as resource allocations are made on the basis of assessing various programs against each other, the matter of problem choice is at least as crucial as quality. However, despite a sizable body of literature on decisionmaking, there are no sure-fire methods for selecting problems or for deciding whether those of highest priority are being addressed. But again, asking some specific questions will help clarify what information is relevant to such an assessment.

- Who thinks the problem or question is important?
- Why is it considered important?
- Are major policy or funding initiatives regarding the problem anticipated?
- How many individuals does the problem affect?
- What is the nature of the injury or disservice done to the individual or group affected?

- What are the overall societal effects of this injury?

The questions themselves imply some methodologies for developing the needed information: opinion surveys, including the opinions of affected populations; collecting opinions of leaders; recording policy as expressed in major federal and state legislation, proposed and enacted; analysis of data from the census, schools, courts, and other sources; statistical and case studies of affected populations. An improved knowledge base should make possible some judgments on relative importance of problems, though the nature of the judgment will still be influenced by the perspectives of the evaluating groups. But problem importance is only one consideration in choice of problems; the second is concerned with feasibility. A problem may be very important, but knowledge and resource constraints may make it a poor choice for the NIE's R&D program. Any major program initiative (except field-initiated basic research) should be subjected to an examination addressing the following questions:

- What are the components of the problem or question that are appropriately addressed through R&D?
- What components of the problem can be ameliorated through educational intervention?
- Has enough R&D progress been made to make further progress likely? To allow needed development and design of alternative educational systems?
- Is there a base of exemplary practice to serve as focus for research? For development? For directed experimentation?
- Are competent people available and interested in working on the problem?
- What other agencies, federal or non-federal, are working on the problem?
- Are the available financial resources appropriate to the likely effort needed? On the part of the NIE? On the part of other agencies that could be engaged to cooperate on the problem?
- If directions for solutions are found or educational alternatives developed, will they be implementable?

Although these ought to be planning questions, they are also relevant to evaluation, particularly if it is to be useful for improving the NIE's R&D strategies. Answers will not always be available at the time programs are initiated; therefore, it should be part of program operations to develop them as a program proceeds. It is the evaluator's function to assess the validity of problem choice in the light of planning rationale and of progress being made toward improved understanding and design of ameliorating interventions.

The two components of problem choice require very different types of information: the first -- on problem importance -- should be as broadly elicited as possible; the second -- on feasibility -- depends on expert knowledge of the state-of-the-action concerning a problem. Review papers similar to those recommended for assessing the state-of-the-art of a field are appropriate here, but with a different focus: to collect and synthesize information on all activities attempting to develop solutions for the problem. Again, quite apart from their importance in making evaluation of problem choice a more rational activity, state-of-the-action reviews would be an invaluable source of information for researchers and decisionmakers of all sorts. Evaluating the NIE's decisions as to the feasibility of R&D approaches to various problems established as important could be carried out by similar panel methods as suggested for the quality evaluation. One product of such an assessment could be suggested changes of problem choice.

Effectiveness of Program Output

Assuming technical quality and appropriate choice of important problems, what are suitable measures for assessing the results of the NIE investment in R&D in education?

First, any evaluation must clearly focus on the fact that the NIE's mission is research and development,* not large-scale action programs, the more common subject of evaluation. Therefore, relevant

*This does not preclude development of strategies and tools for dissemination and implementation of the results of its R&D efforts; it does preclude wholesale funding of adoptions of innovations.

criteria will measure progress in three areas: (1) contributions to the knowledge base needed to deal with educational problems, (2) contributions to policies that further educational objectives, and (3) development and testing of products and processes designed to improve delivery of education. Aggregate measures such as national or city-wide reading scores, dropout statistics, or distribution of different population groups in institutions of higher education are not appropriate in the first few years of the NIE's existence; they may become so provided that NIE-initiated policies or educational interventions become widely implemented, and that the phenomenon being measured is to a substantial degree subject to modification through education.

Second, while some objective criteria are available for assessing program effectiveness in the three areas noted, efforts to develop benefit-cost ratios for purposes of resource allocations are not likely to be any more productive for the NIE's programs than for other R&D support activities. R&D is a risky activity, as Rivlin (1971, p. 51) comments by way of illustration: "The costs of finding a cure for cancer are inherently uncertain; they depend on unforeseeable outcomes of basic and applied research." Nor will it be either possible or desirable to project benefits solely in economic terms for most contributions to knowledge about educational problems or to educational interventions, though there may be some specific initiatives for which this is appropriate, for example, efforts to increase educational programs designed to make migrant rural families economically viable. In general, there will be few instances in which enough empirical data are available to allow the application of cost-benefit analysis.* However, comparative operational costs of educational alternatives developed under NIE auspices are a legitimate evaluation criterion, as noted below.

Assessing progress in the knowledge base needed to resolve questions or problems in education is closely related to the quality assessment. The state-of-the-art reviews suggested there, if the same field is re-examined at periodic intervals, will serve as evidence of contributions

---

*See Rossi, Chapter 2, in Rossi and Williams (1972).

to crucial data, theory, and conceptual understanding ascribable to NIE-supported activities. In addition, use indicators are appropriate:

- o What is the quality and quantity of literature resulting from NIE support?
- o How frequently are findings cited in later work in the field? By researchers not receiving NIE support?
- o Are advances in the understanding of a specific problem or question clearly discernible over a two-year period? A five-year period?
- o Are the findings useful to the NIE's own programs? What is the level of direct application within the NIE?
- o Are the findings being used by other institutions, federal and local? To what extent?

Insofar as these criteria involve judgments of quality of the R&D output, peer-group review is again an appropriate method; amount of usage should, however, be established independently through such means as citation indexes and can in itself help in quality assessment.

Questions to be asked in evaluating contributions to policy formulation also revolve around usage, but the documentation is likely to be much more difficult, since the basis for most policy decisions is usually multifaceted and not often fully explicated. The user clientele, instead of professionals in various disciplines and in education, will be the components of the executive branch of the federal government concerned with educational policy, Congressional committees dealing with education, state and local education agencies, and educational systems and institutions. The documents to be examined, rather than the scientific and professional literature, should include sponsored and enacted legislation at all levels of government, policy statements, and editorial and similar non-professional literature intended to influence public policy. Unlike the somewhat similar search to establish problem importance, the required examination should -- if possible -- be carried out independent of the NIE funding, since its objectivity is likely to be questioned otherwise.

It may be useful, however, for the NIE to sponsor retrospective studies like TRACES* and Project Hindsight** some five or ten years hence, to analyze use of the NIE output both in the knowledge base and policy formulation areas. The NIH, for example, is currently engaged in some examinations tracing the effects of their past efforts. The purpose of such studies should be to enhance program effectiveness rather than influence resource allocations through justification of past support. Therefore -- unlike the examples just given -- the studies should also note instances of failure, particularly in the policy area; for example, where directions were taken in deliberate contravention to what appeared to be indicated in NIE-developed information, or where such information was ignored because of gaps in communication.

Evaluation of success in developing and testing improved products and alternative systems for education can build on a considerable history of such assessment. Educational innovations may consist of designing components that will help make existing systems work better, such as new curriculum programs, information systems accommodating tracking of individualized instruction, performance-based testing to credit experience-based learning; or it may put a number of components together in such a way that an entire new system results. Each of these should be assessed separately, for it is quite possible that some components may prove successful apart from the system for which they were designed. Indicators of success should be based on operational objectives; decisions as to implementation are also relevant criteria, but use criteria should be applied only after broad-scale implementation has actually been attempted. Again, retrospective studies may help highlight the sources of success and failure in development, testing, and implementation. Appropriate questions are:

- Has the developed product or system had the effect originally aimed for, as documented by testing?

---

*_Technology in Retrospect and Critical Events in Science_ (1968), prepared by IIT Research Institute.

**Office of the Director of Defense Research and Engineering (1969).

- o For what populations, in what settings, does it have the desired effect?
- o In what ways, desired and undesired, is the performance and behavior of participants changed by the educational innovation?
- o Is adequate information being provided on how to install the innovation? On costs? On training prerequisites for staff? On special requirements (e.g., equipment, space, management arrangements)?
- o Have the NIE innovations led to implementation funding by social action agencies such as OE or OEO?
- o Are local school systems or other educational institutions investing their own funds in adopting NIE-sponsored products?
- o What are the barriers to implementation?

If implementation actually does take place, additional criteria can be applied, such as number of users or sites, effectiveness of replication (is the product or process still recognizable after it is out of the hands of the original developers?), test scores and other performance indicators, distribution of use among target populations, and unintended side effects.

Assessment of the products of development and experimentation can in itself become a major R&D activity. Planning for appropriate evaluation should be part of the program development process, as emphasized by Crawford (1972) in his recent study of the impact of educational R&D products, but ordinarily the level of evaluation effort will be minimal at program inception and become greater as products come into use. Putting the matter another way, development of truly innovative educational curricula or practices is complex and time-consuming, impact even slower, therefore evaluation of development and experimentation must have an adequate time frame. Considering the high expectation for visible successes, however, which is likely to enter any outside evaluation of effectiveness, the NIE would be well-advised to invest in some short-term projects that could yield rapid payoff, for example, implementation manuals for adopting improved practices that have already been tested through natural experimentation or through demonstration funded by other agencies.

Distribution of Funds and Second-Order Benefits

This dimension of evaluation is quite different in character from the other three: rather than being concerned with outcome, it focuses on process. In some sense, satisfactory performance along the other three dimensions should make this issue superfluous, but it must be considered separately because of its special interest to Congress. Apart from concerns with substantive contribution and allocation of educational R&D resources to yield optimal results, Congress attaches importance to the "fairness" by which R&D funds, prestige, and access to more subtle benefits (e.g., being part of an "in-group") are distributed. Questions of greatest interest usually involve geographic distribution of funds (and also of eventual benefits to practitioners and consumers), widely accessible opportunity to compete for funding (e.g., dislike of sole-source contracts), and openness of management procedures (e.g., 5 U.S.C. 522, *The Freedom of Information Act*). To some degree, the performer communities will share these interests, though their notions of fair distribution criteria will not match those of Congressional or departmental watchdogs. Williams (1971, p. 135) points out that public agencies have traditionally been sensitive to such questions and will attempt to establish a record of accountability and fiscal prudence, sometimes to the point where "administrative purity may become a public manager's greatest concern."

There will never be an adequate response to distributional questions, however, precisely because "fairness" is perceived differently by different overseers and clients, and because any concept of fairness is to some degree in conflict with quality and effectiveness criteria in the allocation of R&D support. The NIE must put quality and effectiveness first, but it should be open to judgment on the availability of information about any of its practices and rationales for them. This implies the existence of an effective management information system that permits quick access to data on number and origins of proposals; data on location and types of performers working on current grants and contracts; agency guidelines on requests for proposals, proposal evaluation, and property rights and licensing procedures for products developed with NIE support; monitoring procedures, and so forth. As important as forthright and prompt response to questions on the *what* of actual practice is the *why*. Therefore, any evaluation should consider

the validity of the reasons for various management procedures, the clarity with which procedures are explained to all concerned parties, and the effects of the procedures. Evaluation should also consider to what extent practices are designed ahead of time in pursuit of deliberate strategies for R&D management instead of representing the accretion of ad hoc decisions and responses to hostile criticisms that characterizes many government programs.

Assessing R&D Capability

The reader will note that the evaluation criteria and methods discussed so far address in a variety of ways the first three missions of the NIE as delineated in the legislation, but few are directly applicable to the fourth, "building an effective educational research and development system." (Although distributional criteria are sometimes made to serve this purpose, they are no more applicable for gauging the effectiveness of educational R&D than they are for gauging the effectiveness of R&D to develop alternative energy sources, despite the great differences in the spread of expertise in the two areas.) This omission is quite deliberate and derives from appraising past attempts at building R&D capability in vacuo, that is, without an existing core of quality R&D, before important problems amenable to R&D approaches are defined, and in the absence of any strategy for assessing the effectiveness of the R&D system's output.

If the NIE can perform successfully in regard to its first three missions, building R&D capability only as specifically required for program initiatives in regard to those missions, then it will indeed be developing an effective educational R&D system, and this will be evidenced through evaluation addressing the substantive missions. Criteria solely concerned with the R&D system itself, e.g., number of educational researchers trained, number of institutions active in educational research, number of new performers, are, in my opinion, not only irrelevant but misleading, for they may raise unwarranted expectations of performance. Such indicators will not be needed to assess the effectiveness of an R&D system that produces the substantive results sought in the NIE's authorizing legislation regarding problem-solving in education, advancing

its practice, and strengthening scientific and technological foundations; nor will they convince in the absence of substantive results.

III. THE USES OF EVALUATION

In considering the various ways in which the NIE should -- and will -- be evaluated, one must ask two further questions: (1) How useful will any evaluation be? and (2) How will evaluation results be used? While the second depends in part on the first, it also depends on political considerations that need to be examined separately from usefulness, for evaluation "cannot (and should not) replace politics, but it can, over time, facilitate better political decisions" (Williams and Evans, 1969, p. 130).

Usefulness of Evaluation

Any evaluation, to be useful for decisionmaking, must have three characteristics: it must be competent; it must be relevant; and it must be honest. Unfortunately, particularly where evaluation is to provide feedback for improving an agency's R&D strategies, these aims may be in conflict, as has been noted by Glennan (1969).

I have suggested several types of studies that need to be carried on fairly continuously in order to provide a substantive information base for evaluation and increase its caliber. This background work is unlikely to get done on a systematic basis unless the NIE itself sponsors a good portion of it. "Unless legislation or agency policy specifically earmarks funds, evaluation staffs will not be assembled nor the evaluation job done. Only when a flow of resources exists will a formal responsibility to evaluate be translated into significant evaluation activities" (Wholey, et al., 1971, p. 77). Thus, to obtain competent evaluation, agency commitment is necessary.

Wholey also points out that spending program funds on evaluation (often resisted by program managers who may view it as a threat) is justified if program decisions are likely to be influenced by evaluation. Relevance to decisionmaking, particularly within the agency, again requires agency involvement, as has been emphasized by nearly everyone who has examined the field, including several of the authors already cited. But

both competence and honesty require objectivity, and that implies that evaluation should be carried out as an independent activity by outside experts. Perhaps the Advisory Council could play the role of sympathetic but impartial judge, but this precludes its functioning as a knowledgeable advocate of educational R&D, another possible role for the Council. In any case, no matter how the Council defines its functions, its credibility with outsiders as objective assessors of the NIE's performance will not be high, raising the old question: *Quis custodiet ipsos custodes?*

For the NIE's own needs, a possible resolution of the quandary is to emphasize competence and relevance in its self-initiated evaluations. To ensure these and the maximum attainable degree of honesty, a threefold strategy might be used in which the NIE Director and Advisory Council define the purpose of the evaluation, and the NIE funds the necessary background studies, but the actual evaluation procedures are carried out as much as possible by outsiders. The aim would be to provide maximum feedback for the NIE; however, a second purpose might also be served: if the NIE succeeds in obtaining competent evaluations based on relevant information for its own needs, these evaluations may find their way into the assessments generated by independent overseers and critics inside and outside government. It is to be hoped that such an information flow will take place so that completely independent evaluations can take advantage of the evaluative information base established by the NIE, and the NIE in its turn will welcome and use independent appraisals.

Using Evaluation Results

Let us assume for the present that such a climate for using evaluation results will actually exist. How could the results be used? There are three ways in which an agency or its overseers can attempt to introduce improvements based on evaluation feedback: allocating resources differently (both as to overall agency budget and internally, among the agency's programs), changing the management procedures, and reorganization. The four dimensions suggested for evaluation bear directly on resource allocation and on management procedures; changes in organization will usually be a consequence of changed resources and management. For example, an assessment of the technical quality of the R&D, if it includes

the suggested state-of-the-art and peer reviews, will uncover which fields are being overfunded and which are being neglected, in view of their potential contribution to the NIE's missions. Thus, priority judgments become feasible that are independent of proposal or other client pressure and less subject to proportional in(de)crementalism, the usual criteria for budget allocations. Assessments of problem choice, based on the subjective and objective criteria discussed for problem importance and on state-of-the-action reviews, will also be useful in formulating priorities for budget allocations, for the NIE as a whole and for individual programs. The recent assessments of physics and space research already referred to have, in fact, been able to incorporate priority judgments based on alternative budgets and quantitative scoring. The assessment of effectiveness of output may lead to such suggested changes in management strategies as altering the emphasis on different R&D styles (e.g., less basic research, more development), changing the degree of directiveness and program control, designing new ways of soliciting proposals, changing proposal evaluation mechanisms, and adjusting monitoring procedures. Clearly, quality and problem choice assessments should also feed into the consideration of what management changes might be needed to improve performance. The implications for management of distribution questions have already been discussed.

If suggested changes in resource allocation or management procedures are substantial, their implementation may require changes in agency organization. Depending on the degree of reorganization needed, a separate assessment (perhaps two, one done by an inside and one by an outside group) may be useful to determine the most effective organization for administering the new budget and management procedures.

Application of evaluation results requires that:*

- o New policy directions are articulated clearly.
- o The agency is in a position to institute the changes.
- o Staff are capable of carrying them out.
- o Client groups are willing to adjust.

---

*See Williams (1971), Chapter 8.

The last three conditions are most likely to be met when "changes are modest and take place within the context of a particular ideology, operating primarily to improve efficiency.... These are changes that sometimes can be made by administrative fiat without necessarily arousing professional opposition.... [But] change in policy and agency ideology... could be experienced as 'revolutionary' and threatening by many of the existing staff [and clients] and therefore would likely be opposed or subverted. Such major changes might only become acceptable when an agency experienced a crisis or a keenly felt need to re-examine existing practices...extraordinary efforts on the part of leadership, perhaps including the introduction of new personnel, might be necessary" (Glaser and Ross, 1971, p. 54). In the end, whether any changes actually take place as a result of evaluation, whether the status quo is preserved despite indicated directions for improvement, or whether changes take place independent of evaluation results will depend to a large extent on the motives of those individuals or groups responsible for generating the evaluations. The motivation is not often truth for its own sake; as Levine and Williams (1971, p. 31) say: "Ordinarily, however, decisionmakers [or those who wish to influence them] have preconceptions about answers to the questions addressed by an evaluation.... A decisionmaker with strong a priori views...will be a good customer for evaluation only when it supports these views." Further, no evaluation will be so free from flaws that it cannot be used or attacked to serve a particular group's purpose.* Only commitment at top management levels to base agency policy on evidence supplied by evaluation results and to implement suggested changes will make evaluation a useful activity.

Besides attempting to ensure the competency, relevance, honesty, and usefulness of the evaluations and evaluation components that it sponsors itself, can the NIE affect in any way the climate in which it will be evaluated?

---

*Williams (1971, p. 123) states this as "the iron law of absolute evaluation flaws.... *The absolute methodological and logistical deficiencies in any evaluation make political infighting a near certainty when evaluation results threaten a popular program.* In short, 'questionable evaluation practice' can always be attacked on methodological grounds for political and bureaucratic purposes" [italics in original].

I believe it can, through assuring positive results of an evaluation that I have not as yet discussed, but that is probably the most important of all: the reactions to the day-by-day signals broadcast by the management and staff of the NIE in all its operations. Whether dealing with prospective performers and their institutions, with its official overseers in the legislative and executive branches, with education professionals or the consumers of education, or with the press and other media, the NIE will be subject to covert and continuing appraisal. Through their words and actions, the staff will project an image of competence or incompetence; of judgment and taste or mediocrity; of a dynamic and flexible enterprise likely to accomplish something, or another manifestation of government bureaucracy. No matter what the formal evaluation mechanisms set up by the NIE itself or by others to evaluate its performance, they will be permeated by the agency's image as created by the staff. There is no more important concern for the NIE, for its ability to carry out its missions and any judgment on its worth will ultimately depend on it.

BIBLIOGRAPHY

Cain, Glen G., and Robinson G. Hollister, "The Methodology of Evaluating Social Action Programs," in *Discussion Papers*, Institute for Research on Poverty, University of Wisconsin, 1969.

Cohen, David K., *The National Institute of Education: What Can Be Expected?*, National Institute of Education, Planning Unit, Report No. P106, Washington, D.C., September 1972.

Crawford, J. J., et al., *Evaluation of the Impact of Educational Research and Development Products*, American Institute for Research in the Behavioral Sciences, Pittsburgh, 1972.

Education Amendments of 1972, 92d Congress, 2d Session, House of Representatives, U.S. Government Printing Office, Washington, D.C., Report No. 92-1085, May 1972.

Fox, David J., "Issues in Evaluating Programs for Disadvantaged Children," *Urban Review*, Vol. 2, No. 3, December 1967, pp. 5, 7, 9.

Glaser, E. M., and H. L. Ross, *Increasing the Utilization of Applied Research Results*, Human Interaction Research Institute, Los Angeles, March 1971.

Glennan, Thomas K. Jr., *Evaluating Federal Manpower Programs: Notes and Observations*, The Rand Corporation, RM-5743-OEO, September 1969.

Guttentag, Marcia, "Models and Methods in Evaluation Research," *Journal of Theory and Social Behavior*, Vol. 10, No. 1, April 1971, pp. 75-95.

Levine, R. A., and A. P. Williams, Jr., *Making Evaluation Effective: A Guide*, The Rand Corporation, R-788-HEW/CMU, May 1971.

Light, Richard J., and Paul V. Smith, "Choosing a Future: Strategies for Designing and Evaluating New Programs," *Harvard Educational Review*, Vol. 40, No. 1, Winter 1970, pp. 1-28.

Project Hindsight Final Report, Office of the Director of Defense Research and Engineering, Washington, D.C., 1969.

Physics in Perspective: Recommendations and Program Emphases (Bromley Report), National Academy of Sciences, Washington, D.C., 1972 (other volumes forthcoming).

Priorities for Space Research, 1971-1980 (2d printing), National Academy of Sciences, December 1971.

Rivlin, Alice M., *Systematic Thinking for Social Action*, The Brookings Institution, Washington, D.C., 1971.

Rossi, Peter H., and Walter Williams (eds.), *Evaluating Social Programs*, Seminar Press, New York and London, 1972.

Rothschild, Lord, "The Organization and Management of Government R. and D.," in *A Framework for Government Research and Development*, Her Majesty's Stationery Office, London, 1972.

Smith, Marshall S., and Joan S. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review*, Vol. 40, No. 1, Winter 1970, pp. 51-104.

Technology in Retrospect and Critical Events in Science, Illinois Institute of Technology Research Institute, Vol. 1, December 15, 1968.

Timpane, Michael P., "Educational Experimentation in National Social Policy," *Harvard Educational Review*, Vol. 40, No. 4, November 1970.

Wholey, Joseph S., et al., *Federal Evaluation Policy*, The Urban Institute, Washington, D.C., 1971.

Williams, Walter, *Social Policy Research and Analysis: The Experience in the Federal Social Agencies*, American Elsevier Publishing Company, Inc., New York, 1971.

Williams, Walter, and John W. Evans, "The Politics of Evaluation: The Case of Head Start," *The Annals of the American Academy of Political and Social Science*, Vol. 385, September 1969.